Slide 1 of 40

Title Slide: Considerations for Analysis of NAEP Data

Slide 2 of 40

This module provides a demonstration of how to use NAEPEX Software for creating syntax files and AM Software for analyzing NAEP data. It also describes considerations for analyzing NAEP data.

Slide 3 of 40

Many statistical packages assume simple random sampling. NAEP data should not be analyzed using such procedures, as they may result in biased estimates. They can also result in underestimated sampling errors that will produce incorrect p values, which can incorrectly indicate differences that are statistically significant when they are not. Therefore, it is necessary to use special statistical software that accounts for the NAEP study design. While several statistical programs such as WesVar and Stata have this capability, and while SAS and SPSS offer the capability through optional add-ons, this module will focus on teaching you two particular tools —NAEPEX Software and AM Software.

Slide 4 of 40

NAEPEX Software is the data extraction program for choosing variables, extracting data, and generating SPSS, SAS, or Stata control statements necessary for the analysis of NAEP data. It is part of the NAEP Data Toolkit, which can be accessed directly from the NAEP restricted-use CD-ROM, or by clicking on the corresponding underlined screen text. NAEPEX is also NAEP's Electronic Code Book, or ECB. As such, it will take you through the process of selecting variables and producing syntax, or script, to generate the syntax for creating NAEP data files you need for your specific analyses. The NAEPEX screens in this module provide information about the use of NAEPEX to complete an example analysis. The NAEPEX Tutorial and User Guide includes more detailed, step-by-step instructions for the installation and use of the NAEPEX program. This guide is provided as a resource document that can be accessed by clicking on the underlined screen text.

Slide 5 of 40

AM Software is statistical software developed by the American Institutes for Research, or AIR. It is designed for use with large-scale assessments and can be downloaded by clicking the corresponding underlined screen text.

AM uses SPSS, SAS, or Stata syntax to perform statistical analyses with NAEP data files. The software also generates code that appropriately uses plausible values, and takes into account complex sampling design in the computation of sampling variance. AM is the proper software for analyzing NCES data that contain plausible values.

Slide 6 of 40

Once users get the NAEP data CD, both the NAEPEX Software and AM Software will be needed for analyses. The data on the NAEP data CD is arrayed in a flat file (a rectangular matrix). NAEPEX will help extract the variables of interest for analyses from the flat file into a statistical software package (SPSS, SAS, or Stata), by generating the appropriate syntax to create an SPSS, SAS, or Stata dataset. Next, users will edit the data by performing any necessary cleaning and recoding. Then the new dataset can be imported into AM Software. Once the NAEP data has been imported, AM Software will be used to conduct analyses.

In the next few slides we will look at a sample analysis using these procedures and SPSS software.

Slide 7 of 40

Here are three sample questions that we will explore, using NIES 2007 data:

- Is there a difference in 2007 reading achievement among public school 8th grade students based on the student gender and the amount of time the teacher spends on reading skills and strategies?
- What is the proportion distribution for days absent from school by gender for public school 8th grade students in 2007?
- Is there a relationship between 2007 reading achievement in the Midwest region among 8th grade public school students, absenteeism, and the number of books at home, while controlling for student's gender and the percentage of American Indian students in the school?

It might be helpful to either print this list of questions to refer to as we go through the examples in the next few slides, or to click the underlined screen text, **Example** to view and print the corresponding resource document.

Slide 8 of 40

Open the NAEPEX program once it has been installed on your computer. If no data sources are detected on the system, the first window that appears will be as shown on the screen here - the Currently Available Data Sources screen is blank. To begin, click on **Add New** in the Data Sources Window.

Slide 9 of 40

Next, in the **Browse For Folder** window, navigate to the appropriate location, where the data are stored. If the data source is a NAEP restricted-use data CD/DVD, enter the NCES-provided password to unzip the data file to a location on your hard drive first. For this example, the data source location is C:\NIES Seminar\Data\NIES2007. Once the main folder of the data source has been located, click OK to copy it to the NAEPEX repository. It will take a few minutes for the data to be loaded. Once the loading is

completed, click **OK** to go to the Data Sources window where the data source will now appear.

Slide 10 of 40

Once the data source appears in the Data Sources window, select the data source and click **Open**. After opening the data source, select the data and click **Open Data File**. For the example analysis, the data source is **2007 National Indian Education Study (NIES) parts I and II grades 4 and 8** and the data is **National Reading Assessment: Grade 8 Student, Teacher & School data**.

Slide 11 of 40

Next, a Selection Information screen appears and the required information must be filled in before the additional tabs (i.e., Select Variables, Filter Data, and Generate Files) become available. Users will need to type a Selection Description into the appropriate box. The description is text that NAEPEX writes at the beginning of each file later generated. This allows users to write something that helps in identifying the analysis in the future. It is not necessary to include the entire data set name since this will automatically appear in the output. NAEPEX provides a default file name that can be modified. Here, we have typed the description **Selection 1 for AM demo** and the file name **NAT8R07**. Placing a checkmark in the School Data Included checkbox allows NAEPEX to correctly match students to schools when using both school-level and student-level variables and to produce a single merged file of variables. If the box is left unchecked, there would not be an opportunity to select school-level variables for student-level analyses. In addition, placing a checkmark in the Preselected Data Used checkbox causes NAEPEX to automatically include a number of NAEP variables that are commonly required for analyses. For example, the student replicate weights are almost always selected in order to conduct an analysis properly. If left unchecked, these essential variables must be chosen later in the process.

Slide 12 of 40

Here is the list of variables that NAEPEX pre-selects:

- RRPCM1-RRPCM5 (spelled R-R-P-C-M-1 and R-R-P-C-M-5): Plausible values, NAEP reading assessment
- ORIGWT (spelled O-R-I-G-W-T): Student weight
- JKUNIT: Jackknife variance unit
- REPGRP1 (spelled R-E-P-G-R-P-1): Jackknife variance stratum
- SRWT01-SRWT62 (spelled S-R-W-T-0-1 AND S-R-W-T-6-2): Replicate weights
- CENSREG: Census region of the country
- DSEX: Gender

- FIPS: FIPS state code

Slide 13 of 40

The Select Variables tab has a choice of views for available and pre-selected variables, as well as a search feature to help users find the variables of interest. To search for a variable, type a key-word in the Search String box and click the Search button. To answer the questions in the sample analysis, a variable describing the number of books at home is needed. Here, we have searched for the key word **Books**, which produced a window showing that 11 records were found that match **Books**. The **Yes** under the **Has Details** column means that further information about the variable, such as data values, labels, and frequency counts, is available by right-clicking on the variable name. We have clicked the checkbox for the variable **B013801** to choose the variable for **Books home**. After clicking the Select button, this variable will appear under the View Selected tab. Following the same procedures, we will select the variables **B018101**, **Days absent from school last month**, and **T083601**, **Percent of time spent on reading skills and strategies** for the example analysis.

Slide 14 of 40

We also searched for and selected the variables for school type and percent American Indian for the example analysis. Note that the variables for gender (DSEX) and geographic region (CENSREG) that are needed for the example analysis, were two of the variables pre-selected by NAEPEX, so we did not need to select them. The reporting sample variable (RPTSAMP) is the flag that indicates a student took the NAEP assessment, so this variable must be selected for the analysis.

It is very important to select ALL the variables needed for the analysis before creating the SPSS or SAS syntax. Otherwise, users will need to return to NAEPEX and start the procedures again. The NAEPEX Tutorial and User Guide resource provides more information about selecting and deselecting variables, as well as the choices for viewing variables.

Slide 15 of 40

After you have selected all of the variables for your analysis, click on the **Generate Files** tab and then check the box for your choice of the type of syntax file (or files) to generate. SPSS is used for the example analysis. Click on the **Generate Files** button and then click **OK** to save the changes to the file you created. A pop-up window will appear to indicate the syntax files generated and their locations and you will need to click the **OK** button again. Once you have generated the syntax file (or files), close NAEPEX.

Slide 16 of 40

Next, open your SPSS program to run the syntax file and generate an SPSS dataset. On the SPSS Menu Bar, go to **File**, select **Open**, and then select **Syntax**. Navigate to the location where you saved the SPSS syntax file generated by NAEPEX. Select the syntax file and then click **Open**.

Slide 17 of 40

From the menu bar, select **Run-ALL**. To save the file as an SPSS data file, go to **File** on the SPSS Data Editor, then choose the **Save as** option to save the file as a **.sav** file. Note the file name and location, and then close SPSS.

Slide 18 of 40

Now, open the AM Software. When the introductory screen appears, click on the screen to start. On the Menu Bar, go to **File**, select **Import**, and then select **SPSS.sav File**.

Slide 19 of 40

When the dialog box appears, provide the path and file name of the SPSS data file you saved in the last step and click **Open**.

Slide 20 of 40

Next, save the SPSS data file as an AM data file by selecting **File**, then **Save As**, and then typing a file name. Be sure to save the file as a **.am** file.

Slide 21 of 40

Next, users will need to define the Weight, Cluster and Strata to reflect the NAEP sample design. To do this, right-click on the weight variable and then click on **Edit Metadata**. The weight variable in the example data file is ORIGWT. Then, in the dialogue box, check the appropriate **Design Role** button, which is **Weight**. Then click **OK**. To set the cluster variable, right click on the cluster variable and then click **Edit Metadata**. The cluster variable in the example data file is JKUNIT. Next, in the dialogue box, check the **Cluster** button and then click **OK**. Similarly, right-click on the strata variable, which is **REPGRP1** in the example data file, and click **Edit Metadata**. Then, in the dialogue box check the **Strata** button and click **OK**.

Slide 22 of 40

Next users will identify the replicate weights. First, right-click on any variable, and click on **Set Replicate Weights** from the menu. Highlight all the replicate weight variables and drag them into the Replicate Weights box of the Set Replicate Weights window. The replicate weights in the sample data file are SRWT01 (spelled S-R-W-T-0-1) through SRWT62 (spelled S-R-W-T-6-2). Now, set the replication method to JK2 by clicking the appropriate button and then click **OK**. It is important to note that users should always select the replication method that is appropriate to the dataset of interest, which for NAEP is JK2.

Slide 23 of 40

At this point, the data are ready to be edited. The first step is to define the missing values for the variables that have been selected for analysis. Information about the variables of interest for the analysis is included in the Electronic Codebook, or ECB. For the example that follows, we have used information about three variables described in

the sample analysis from the ECB, and provided it in a table. Those variables are B013801, B018101, and T083601. The table shows that for each variable, a response value of 0 indicates that the respondent provided multiple responses to the question, while a response value of 8 indicates that the respondent omitted the question. The numbers under the variable name indicate the total number of respondents for which these conditions apply. For example, 1 respondent provided multiple responses to question B013801; while 601 respondents omitted the question pertaining to variable B013801. This variable, B013801, corresponds to a question regarding the number of books at home. The values of 0 and 8 must be set to missing so that the respondents with these values will not be included in the analyses.

Slide 24 of 40

To set values to **missing** within AM for B013801, right-click on the variable name and choose **Set values to missing**. The dialog box that opens presents some basic descriptive information about the variable's values. Using the radio buttons, indicate whether you want to set all values equal to, greater than, less than, or, not equal to a value. Next, specify the value in the edit box below and click **OK**. A pop-up window will indicate the number of values that have been changed to **missing**. The screen displayed here shows that for variable **B013801**, **Number of books at home**, the value of **8** is set to missing. For the sample analysis, we also set the value of **0** to missing, and repeated the procedures for the variables **B018101**, **Number of days absent from school last month** and **T083601**, **Percent of time spent on reading skills and strategies**. Note that this operation modifies the original variables.

Slide 25 of 40

Depending on the analyses being run, users may want to collapse categories of a categorical variable. Researchers can either modify the original variable by selecting the checkbox in the top left of the dialog box; or, create a new variable. For the sample analysis, we wish to recode the variable **T083601**, **Time spent on reading skills and strategies** as a new variable. To collapse response categories, first, right-click the variable name and choose **collapse categories**. Next, type a new variable name for the recoded variable. By default, AM will automatically generate value labels to remind you of the mapping between the original variable and the new coding. However, you can de-select this option and type your own variable labels. In the example shown, we have named the new variable **Readtim** (pronounced “read time”; spelled R-E-A-D-T-I-M) and let AM auto-generate a value label. In the list at the bottom part of the dialog box, two columns of codes appear: the original code, and a new default recoding. The new coding appears in the left-most column. To change codes, click on the new code and enter the code that you would like to replace it with. If you associate the same new code with two or more original codes, these categories will be collapsed together. In the example, the values of 1 and 2 have been collapsed and recoded as “1”; the values 3 and 4 have been collapsed and recoded as “2”; and, the values 5 and 6 have been collapsed and recoded as “3.” To change a value to **missing**, click on the new code, and then click somewhere else without entering a new code. Once you have re-coded the variables click **OK**.

Slide 26 of 40

"Dummy coding" a variable creates a set of indicator variables to represent the information in a categorical variable. AM Software generates a series of dummy variables for each original variable by adding a number starting with 0 to the end of the base variable name. For our sample analysis, we dummy coded the DSEX variable, which has two values: 1 indicating male, and 2 indicating female. Dummy coding produced two variables: DSEX0, which represents males and DSEX1, which represents females.

Slide 27 of 40

To dummy code a variable, right-click on the variable name and then click on **dummy code this variable**. Enter a new variable name in the **Base name for new variables** window, and then click **OK**. Here, the variable **B013801**, **Number of books at home** is dummy coded with the new base name **BOOKS**. As the variable B013801 has four categories, scroll down to the bottom of the variable list to view the four new dummy variables, BOOKS0 through BOOKS3.

Slide 28 of 40

For the example analysis, the variable **B018101**, **Days absent from school last month** is also dummy coded with the base name **ABSENT**. Scroll down to the bottom of the variable list to view the five new dummy variables, ABSENT0 through ABSENT4.

Slide 29 of 40

To analyze only a subset of the data, a filter will need to be set. To set the filter, go to **DATA** and then select **FILTER OBSERVATIONS** to create an expression that filters the data used in the analysis.

The sample analysis concerns public school 8th grade students who were in the reporting sample. To filter only this subset, we type an open parenthesis into the Selection Criteria window. Next, we select the variable **RPTSAMP**, then the **= operation**. These selections will appear in the Selection Criteria box. Then we type a **1** into the Selection Criteria box to designate that only those students in the reporting sample be included in the analysis and finish the expression with a close parenthesis.

Next we select the asterisk operation and type another open parenthesis to begin the second part of the filter. We select the variable **SCHTYPE** (pronounced "school type"; spelled S-C-H-T-Y-P-E) and the **= operation**, and then type a **1** into the Selection Criteria box. This designates that only public school students be included in the analysis.

Once filter selections have been made, click **OK**.

Slide 30 of 40

Now, we are ready to answer questions in the sample analysis.

To answer the first question, “Is there a difference in 2007 reading achievement among public school 8th grade students based on the student gender and the amount of time the teacher spends on reading skills and strategies?” it is necessary to run a cross-tabulation, or cross-tab, of the mean scores.

Go to **Statistics**, in the main menu, select **Plausible Values Procedures** and then select **Means**. In the window that appears, type a title for the crosstab.

Next, drag the dependent variables into the **Dependent Variables** box and the Independent Variables into the **Independent Variables** box. Here, the dependent variables are **RRPCM1** through **RRPCM5**, which are the reading achievement score plausible values. The independent variables are **GENDER** and **READTIM**.

Finally, click **OK**.

Slide 31 of 40

When you click **OK**, the results open in a new browser window, or output in a spreadsheet or text format, depending on the output option you choose. The results include the weighted N, the mean, the standard error, and the standard deviation.

On the screen we can see the results from the sample crosstab analysis.

Slide 32 of 40

To check the statistical significance of the difference in the means in your analysis, you will need to conduct a t-test. In the Completed Runs section on the right side of the AM Software window, right click **PV Means Plausible Value Means**, and then click **T-Tes**t. When the **Perform T-Tests** window opens, click on the mean of the given group, and then click on the mean of a contrast group. The t-test appears at the bottom of the screen, and a result highlighted in yellow indicates that the difference is significant. Here, the difference between males and females not absent, or absent, 1-10% of the time was not significant at the 0.05 level of significance. The difference between males and females absent either 11-40% or 41-60% of the time is significant at the 0.05 level of significance. After you have added your t-tests, click **OK**.

Slide 33 of 40

Similar to the Crosstab results, when you click **OK**, the t-test results are opened in a new browser window, or output in a spreadsheet or text format. Here are the results from the sample t-tests.

Slide 34 of 40

To answer the second question of the example analysis, “What is the proportion distribution for days absent from school by gender for public school 8th grade students in 2007?” it is necessary to use the frequencies procedure.

In the main menu, go to **Statistics**, select **Replication Procedures for Basic Statistic**, and then select **Frequencies**. In the window that appears, type a title for the frequency analysis.

As with the Crosstab procedure, drag the dependent variable into the **Row Variable** window and the independent variable into the **Column Variable** window. Here, the variable **B018101**, **Days absent from school last month**, is the dependent variable and **DSEX**, **Gender**, is the independent variable. The appropriate replicate weights will auto fill.

Once you have selected the variables for the frequency analysis, click **OK**.

Within the AM software package, it is important to take care when dragging dependent and independent variables into different model specification windows.

Slide 35 of 40

The results of the frequency analysis include the weighted N, the frequencies, and the standard errors.

Slide 36 of 40

To answer the third question of the example analysis, “Is there a relationship between 2007 reading achievement in the Midwest region among 8th grade public school students, absenteeism, and the number of books at home, while controlling for student’s gender and the percentage of American Indian students in the school?” it is necessary to run a regression analysis.

Because this analysis is based on 8th grade students in the Midwest region who were in the reporting sample, it is first necessary to set a filter. Using the procedures covered previously, we set the filter for **RPTSAMP=1** (to include only those students in the reporting sample), and **CENSREG=2** (to include only those students in the Midwest region).

Slide 37 of 40

To run the regression analysis, in the main menu, go to **Statistics**, then select **Plausible Values Procedure** and then **Regression**. In the window that appears, type a title for the regression analysis. As with the previous procedures covered, drag the dependent variables into the **Dependent Variables** box and drag the independent variables into the **Independent Variables** box. Here, the dependent variables are the reading achievement score plausible values, **RRPCM1** through **RRPCM5**. The independent variables are **GENDER**, **BOOKS1** through **BOOKS3**, **ABSENT1** through

ABSENT4, and **PCTIND** (pronounced “percent American Indian”). Finally, click “**OK**” to start the procedure.

Slide 38 of 40

The results of the regression analysis include the estimate, the standard error, the z score, and the p-value.

Slide 39 of 40

More detail on AM Software is available in the AM online manual and help system, which can be accessed on the AM website by clicking on the corresponding underlined screen text.

Slide 40 of 40

This module has described the use of NAEPEX Software and AM Software. It has also described the analytic considerations that should be kept in mind when using NAEP data. The module’s objectives are summarized here for your reference. The resources provided throughout the module are also listed and can be accessed by clicking on the corresponding links. You have now completed this series of modules on NAEP. Click the **Exit** button to return to the landing page.